Pluralized non-count nouns across Englishes: a corpus-linguistic approach to dialect typology

Schneider, Gerold ; Hundt, Marianne ; Schreier, Daniel

Abstract: This corpus-based study of pluralized non-count nouns (informations, advices, etc.) uses collocation-derived measures (determiners vs. bare noun and mass quantifiers) to extract potential candidates of non-count nouns in a bottom-up approach from the British National Corpus (BNC), allowing the detection of grammatical categories from distributional features. We then use this token list to retrieve data on pluralization of non-counts from nine annotated components of the International Corpus of English (ICE). While the distinction between count and non-count nouns is gradient rather than categorical, it is still possible to distinguish between standard and non-standard pluralization of non-counts. Qualitative analyses of our data show that non-standard pluralization of non-count nouns is regularly attested in second-language varieties, including previously unrecorded types; however, it is also occasionally found in first-language varieties. We discuss implications of our corpus results for common explanations of pluralized non-count nouns, such as substrate influence, language learning effects and historical input. By combining a bottom-up corpus-based approach with fine-grained qualitative analyses we can provide a more nuanced view of pluralization of non-counts across ENL and ESL for the investigation of World Englishes.

DOI: https://doi.org/10.1515/cllt-2018-0068

Originally published at:
Schneider, Gerold; Hundt, Marianne; Schreier, Daniel (2020). Pluralized non-count nouns across Englishes: a corpus-linguistic approach to dialect typology. Corpus Linguistics and Linguistic Theory, 16(3):515-546.
DOI: https://doi.org/10.1515/cllt-2018-0068

Gerold Schneider*, Marianne Hundt and Daniel Schreier

Pluralized non-count nouns across Englishes: A corpus-linguistic approach to variety types

https://doi.org/10.1515/cllt-2018-0068

Abstract: This corpus-based study of pluralized non-count nouns (*informations*, *advices*, etc.) uses collocation-derived measures (determiners vs. bare noun and mass quantifiers) to extract potential candidates of non-count nouns in a bottom-up approach from the *British National Corpus (BNC),* allowing the detection of grammatical categories from distributional features. We then use this token list to retrieve data on pluralization of non-counts from nine annotated components of the *International Corpus of English (ICE)*. While the distinction between count and non-count nouns is gradient rather than categorical, it is still possible to distinguish between standard and non-standard pluralization of non-counts. Qualitative analyses of our data show that non-standard pluralization of non-count nouns is regularly attested in second-language varieties, including previously unrecorded types; however, it is also occasionally found in first-language varieties. We discuss implications of our corpus results for common explanations of pluralized non-count nouns, such as substrate influence, language learning effects and historical input. By combining a bottom-up corpus-based approach with fine-grained qualitative analyses we can provide a more nuanced view of pluralization of non-counts across ENL and ESL for the investigation of World Englishes.

Keywords: Pluralization, non-count nouns, World Englishes (WE), corpus-based approach, distributional grammar

1 Introduction

Kachru's (1985) model of WE groups countries into three concentric circles: the *Inner Circle* where English is the first or "native" language (ENL) of the majority

***Corresponding author: Gerold Schneider,** Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland; English Department, University of Zurich, Zurich, Switzerland, E-mail: gschneid@es.uzh.ch
https://orcid.org/0000-0002-1905-6237
Marianne Hundt: E-mail: m.hundt@es.uzh.ch, **Daniel Schreier:** E-mail: schreier@es.uzh.ch, English Department, University of Zurich, Zurich, Switzerland

of speakers (e.g. GB or Australia), the *Outer Circle* where English is an institutionalized second language (ESL; e.g. India or Nigeria), and the *Expanding Circle* where English is widely used as a foreign language (EFL; e.g. much of continental Europe). The taxonomic problems of this and related models have been discussed extensively (Schneider 2015). One of them is that the circles model assumes a categorical distinction between variety types. Our paper contributes to the critical empirical investigation of this assumption. The key question is whether there are truly diagnostic features, i.e. features exclusively found in ESL. One of the most frequently cited candidates is the extended use of *–s* pluralization with all noun types (see Section 2.2). Corpus-based research into this issue is still limited and, with the notable exception of Schmidtke and Kuperman (2017), studies have been limited to a predefined set of items. Our methodological goal is to collect an extensive list of pluralized non-count nouns in a bottom-up approach (rather than the few prototypical types usually cited in the literature). We use this list to assess their overall text frequency and use across different WE and discuss their relevance for common explanations such as substrate influence, language learning effects and historical input.

We first provide background information on the distinction between count and non-count nouns, previous corpus-based research on pluralized non-counts and extension of plural marking in WE (Section 2). We describe the corpora that we use as our source of evidence and the procedure to semi-automatically retrieve pluralized non-count nouns (Section 3). The results are reported and evaluated in Section 4, and discussed in the context of variety types in Section 5.

2 Theoretical background and previous research

2.1 Count vs. non-count nouns

Count nouns refer to entities that speakers conceptualize as countable; for those that are not, grammars (e.g. Quirk et al. 1985: 247) distinguish between concrete (*butter*) and abstract (*laziness*) non-counts. On closer inspection, the dichotomy turns out to be too simplistic. Nouns refer to entities whose atomicity appears to range from clearly bounded/solid to amorphous/non-solid (e.g. Jackendoff 1991). In line with this underlying cognitive gradient, the grammatical encoding is gradient, too:

> There are nouns that arguably can be treated as either mass or count (e.g. *bread*). Furthermore, nouns that seem to belong to one class may be coerced to the other by

> specific syntactic constructions. Mass nouns may occur as count nouns; for example *three beers* 'three glasses of beer', *three oils* 'three kinds of oil.' And count nouns may occur as mass nouns; for example *apple* in *Put more apple into the salad!* ... The meaning of a noun occurrence, consequently, is a function of its lexical meaning and the syntactic context in which it appears. (Krifka 1999: 221)[1]

Krifka (ibid.) argues that, in terms of countability, non-count nouns include both "stuff nouns" like *oil, gold, flour* (prototypical mass nouns) and "collective nouns" such as *furniture, cattle, staff*. Grimm and Levin (2011, 2012) distinguish a sub-group of the latter, which they refer to as "functional collectives" (i.e. *furniture, luggage, jewelry*) that "straddle" the traditional mass-count distinction. While count nouns may be coerced into a mass-noun use and vice versa, Cruse (1999: 270) notes that " ... one usage is intuitively felt to be more basic than the other: ... *apple* is basically a count noun and *beer* a mass noun". Non-prototypical uses of stuff nouns, according to him, fall into two groups: 1) a "kind of" reading (e.g. *Chinese green teas*), and 2) a parcelling-out of mass into measurable quantities (e.g. *three beers, two ice-creams*). Moreover, some nouns are hybrid: *These oats are not suitable for muesli* (count) vs. *How much oats have you got in that sack* (mass) (Cruse 1999: 269). Furthermore, polysemous nouns like *control* have both an abstract, non-count reading and a concrete, count-noun sense as in *the controls of an airplane*.

The distinction between count and non-count nouns is further complicated by their morphology: non-count nouns like *measles* are only expected in their plural form, while count nouns like *fish* do not typically inflect for plural. Words like *cattle* are plural in meaning but singular in form, so overt countability (in standard ENL grammar) is only possible with the help of a classifier, i.e. *twelve head of cattle*, whereas words like *scissors* are singular in meaning but plural in form, so a classifier is needed as an additional site for plural marking, i.e. *three pairs of scissors*. Plural morphology on the noun itself is thus not *strictu sensu* a reliable indicator of the count-mass distinction. This poses a methodological challenge for our study (see Section 3).

2.2 Corpus-based approaches to the count vs. non-count distinction

Distributional Information, i.e. statistical analyses of the context, is an important source for deriving word category and syntactic structure, both in theoretical

1 See also Allan (1980) or Joosten (2003). Ziegeler (2010) challenges the coercion view and proposes derivation by metaphorical processes, instead.

linguistics (Harris 1954), computational linguistics (Klein and Manning 2001) and psycholinguistics (Tomasello 2000; Mintz et al. 2014). In this spirit, we focus on previous research that has used bottom-up approaches to the retrieval of non-counts in English. Baldwin and Bond (2003) use a rich set of 1284 features to predict (non-)countability. The features include frequency, several Bayesian probabilities each of number of head and modifier(s), number disagreement in noun conjunctions, absence of determiner, participation in *of* constructions and context-based features such as pronouns and verb number in the vicinity. They achieve an F-score (i.e. the harmonic mean of precision and recall) of up to 89% on assigning the class *uncountable*[2] but do not provide an evaluation of the predictive power of individual features. Moreover, their data come from standard ENL corpus material, only.

Schmidtke and Kuperman (2017) use data from the *Global Web-based English* (GloWbE) corpus[3] and a combination of a bottom-up approach for data retrieval with a top-down approach for variety clustering to study pluralization of mass nouns across ENL and ESL varieties (on the basis of an a priori grouping of varieties). They use a frequency-based approach that is "blind to the count-mass distinction in the initial step" (2017: 141). Moreover, they do not provide a qualitative analysis of their data with respect to the semantics of the pluralized non-counts, i.e. they compare overall pluralization rates and do not distinguish between count-noun coercion and "proper" overextension. Thus, their "coarse-grained", quantitative approach needs supplementing with qualitative analyses, as they (2017: 159) point out themselves.

2.3 Pluralized non-count nouns in WE[4]

The assessment for 76 English varieties in the *electronic World Atlas of Varieties of English* (*eWAVE*, Kortmann and Lunkenheimer 2011), which is based on linguists' rating of vernacular features in WE, yields a clear pattern: pluralization of non-count nouns is particularly frequent in ESL varieties. An A ("pervasive or obligatory") or B rating ("neither pervasive nor extremely rare") is given for 15 out of 18 L2 Englishes (83.3%). Generalized *–s* pluralization has been

2 Their overall F-score is up to 94%, but the rarer class *uncountable* is harder to predict than the dominating class *countable*.

3 See Davies and Fuchs (2014) for a critical discussion of the advantages and limitations of this resource, see the contributions in *English World-Wide* 36(1).

4 This overview focuses on ESL varieties of English. The distinction between count and non-count nouns is also blurred in other contact varieties of English, such as Aboriginal English (see Kortmann and Lunkenheimer 2012 for details).

noted to be particularly common in ESL of Africa and Asia such as Kenya, Cameroon, Indi, Sri Lanka or Hong Kong.[5] According to Mesthrie and Bhatt (2008: 53) "[a]lmost every study of individual WE varieties in Africa and Asia reports frequent examples like *furnitures, equipments, staffs, fruits*, *accommodations*, and less common ones like *offsprings, underwears, paraphernalias*, etc". The feature is absent in the majority of regional varieties of AmE and BrE (8 out of 10). Table 1 lists the ratings that are relevant in our context for the varieties available in the ICE corpora (see Section 3), including vernacular varieties in contact with the non-ENL varieties which we study.[6]

Table 1: *eWAVE* ratings for extended pluralization.

	Rating			
Variety type according to *eWAVE*	A (feature pervasive or obligatory)	B (feature neither pervasive nor extremely rare)	C (feature exists, but extremely rare)	D – (attested absence or other rating)
High-contact L1		Singapore E, Philippine E		American E, Irish E, New Zealand E
Indigenized L2 varieties	Hong Kong E, Indian E	Jamaican E		
English-based Creoles			Jamaican Creole	

Mair (2017: 16) provides evidence from his *Corpus of Cyber-Nigerian* for pluralization of *stuffs*, showing that the nativized pattern is more frequent on webpages in Nigeria than by expatriate Nigerians in the US and Great Britain (see also Alo and Mesthrie 2004: 821), indicating that the feature is susceptible to standardization in dialect contact situations. Mohr's (2016) top-down study of 22 nouns in ICE and GloWbE shows that individual varieties in East Africa differ significantly with respect to the frequency of overgeneralized plural non-counts, thus qualifying the rater-based *eWAVE* description but supporting a general ENL-ESL divide. Schmidtke and Kuperman's (2017) results indicate that pluralized non-count nouns are, indeed, more regularly attested in ESL data; with respect to relative magnitude of pluralization. Moreover, ESL varieties cluster regionally, with e.g. South Asian varieties (Pakistan, Sri Lankan and Indian

5 For details, see Sharma (2012: 525), Wong May (2012: 553), Sand (2012: 212), Mesthrie (2012: 497), Taiwo (2012: 411) and Schmied (2012: 460), Schmied (2008: 454).

6 See https://ewave-atlas.org/parameters/48#2/7.0/7.7 for the complete list.

English) showing a similar propensity for pluralization. An important caveat with respect to Mohr's (2016) and Schmidtke and Kuperman's (2017) results is that they provide frequencies but no qualitative analyses of the nouns in context, i.e. no information on the proportion of regular and non-standard pluralization of non-counts.

Mesthrie and Bhatt (2008: 161) attribute the extension of plural marking in L2 Englishes to learning strategies and transfer. Sharma (2012: 524) adds historical source dialects as a third factor. With respect to historical explanations, it is important to note that pluralization was originally possible for many nouns but then lost from the ENL varieties that served as the original input (see also Denison 1998: 96–98). Examples would be *per cent*, which had a count-noun sense referring to stocks paying a specific interest rate (see OED, s.v. *per cent*, n.) and *advice* in the sense of "opinion", which the OED (s.v. 2.b.) describes as "[n]ow chiefly Caribbean and S. Asian". A cursory glance at historical data (e.g. from the court proceedings of the *Old Bailey* and the *Corpus of Historical American English*) shows that pluralized non-counts are, in fact, regularly attested in earlier stages of BrE ((1)–(4)) and AmE ((5)–(8)):

(1) Q. Had you other *furnitures* of the same kind made up? (OBC, t-**1829**-0115-141)
(2) ... these *evidences* being considered, the Jury brought him in not guilty. (OBC, t-**1675**-0707-8)
(3) by neglecting to provide her with other proper *accommodations*, and forcing her to lay in a damp, wet, and unwholesome cellar, ... (OBC, t-**1784**-0225-63)
(4) Mrs. Cope desired me to go out after a man she saw go out with some *stuffs*; I overtook the prisoner in Eagle street with two pieces of stuff under his arm; (OBC, t-**1803**-0420-13)
(5) you are bound to decide by the *evidences*, the glorious privilege of trial by jury. (COHA, 1820, MAG)
(6) We shall then have six colleges, and twenty-five instructors, and *accommodations* for six hundred pupils. (COHA, 1827, MAG)
(7) we must make our way through some previous *researches*. (COHA, 1824, NF)
(8) if England should withdraw this monitory advice, and again admit our bread *stuffs*, provisions and raw materials. (COHA, 1827, NF)

Regarding the very commonly observed possibility of substrate influence, we have to consider that, while languages universally have ways of expressing the distinction between singular and non-singular (including categories such as dual and plural), not all languages mark number or "numerosity" (Cruse 1999: 267) morphologically in the noun phrase. Wong (2012: 552–553) points out that

> [t]he break-down of mass/count noun distinctions in [Hong Kong English] ... can also be traced back to the syntax of the substrate. The overall structure of a noun phrase in Cantonese is similar to the English one, with the difference that a classifier (CL) is required in the former but not in the latter.

The use of a classifier means that the default class is unclear and can be overridden fully productively in Cantonese.[7]

This fact deserves special attention in the analysis of Englishes that are embedded into high-contact scenarios alongside typologically very different languages. Moreover, semantic and pragmatic aspects play a role when there is a lexical element in cross-linguistic variability: as Cruse (1999: 270) points out, even if languages mark number morphologically in the noun phrase and distinguish between count and non-count nouns (including both stuff and collective nouns), there may be variation in the conceptualization of individual nouns: " ... *spaghetti* is a singular mass noun in English, but plural in Italian and French, ... ; *fruit* is basically a mass noun in English (*Have some fruit*), but a count noun in French ... ". A list of examples of nouns that are "non-count" in English but "count" in other languages (including typologically related languages like German, for instance) includes *accommodation, advice, baggage, equipment, food, homework, information, hair, luggage, machinery, money, news, progress,* and *trouble*.

In second language acquisition, transfer from the substrate language may play a role, but in addition, vernacular features may arise from general mechanisms of language acquisition in contact-induced processes of language shift, such as analogical extension or overgeneralisation, "economy of production" (leading to simplification) and "hyperclarity" (resulting in redundant marking, see Mesthrie 2017: 186–187). However, beyond the outer circle, Hall et al. (2013: 20) show that non-standard pluralization is very infrequent in ELF contexts, concluding that the feature is not helpful in distinguishing ENL from "non-native" varieties, generally. As inner circle and expanding circle are similar for this feature, it can add a new, unexpected pattern to the study of the gradience from ENL to ESL and EFL (Mukherjee and Hundt 2011; Deshors et al. 2016; Schneider and Gilquin 2016; Meriläinen and Paulasto 2017).

The aim of our paper is to test the hypothesis that the extension of pluralization to non-count nouns beyond standard instances of coercion as in *two coffees* is particularly frequent in, and limited to, ESL varieties. We do this using a two-pronged approach: a corpus-based bottom-up retrieval of a list of candidates of non-count nouns (instead of the widely used top-down approach); this

7 A mirror image of this can be found in SingE, where speakers extend the use of bare nouns to count nouns (as in *I have car*, see Ziegeler 2010).

list is then used to retrieve data on potential pluralized non-counts from corpora of WE. In a final step, we analyse our candidates for extended pluralization qualitatively, thus moving beyond Schmidtke and Kuperman (2017) purely quantitative approach.

3 Data and methodology

We use the BNC and the ICE, which were automatically annotated using a syntactic dependency parser (Schneider 2008). ENL[8] data come from ICE-GB (Great Britain), ICE-IRE (Ireland), ICE-CAN (Canada), ICE-NZ (New Zealand) and ESL data from ICE-SIN (Singapore), ICE-HK (Hong Kong), ICE-IND (India), ICE-PHI (Philippines) and ICE-JAM (Jamaica).[9] Like Schmidtke and Kuperman (2017), we apply a bottom-up approach to extract potential pluralized non-counts. While their study relies exclusively on morphological marking and does not distinguish between count and non-count for the retrieval, our approach is more theory-informed in that it uses collocation statistics and morphosyntactic criteria typical of non-counts (see 3.1). The initial results are evaluated and fine-tuned in two steps (3.2 and 3.3). The list obtained from the BNC is used in a top-down approach to retrieve potential pluralized non-counts from the ICE components.

3.1 Semi-automatic retrieval with morphosyntactically motivated collocation measures

According to Krifka (1999: 221), mass nouns (both what he calls "stuff nouns" and "collective nouns") are characterized by three properties:

i) They do not co-occur with the indefinite article *a(n)*: **an oil* but are typically used as bare NPs (and without overt number marking);
ii) they typically do not combine with "number words" (**one cheese*, **three golds*) but can be used in "numerative constructions" (e.g. *five gallons of gas*);
iii) they generally do not co-occur with certain quantificational determiners (**every, many, all oil/butter/chocolate*) in their (default) mass interpretation, selecting a different set of quantifiers instead (*much, little, some, a lot of, a huge amount of oil/butter/gold*).

8 Strictly speaking, JamE is a (standardizing) second dialect (ESD) variety used alongside an English-based creole. It shares with other ESL varieties that it is a high-contact standard(izing) variety of English.

9 The ICE components used in this study are the ones that were available in a parsed format at the time when the data were retrieved. For a critical appraisal of comparability, see Hundt (2015).

In the following, we will illustrate how we operationalized these properties to retrieve potential non-counts.

Property 1: The use of bare NPs (e.g. *I like Ø milk*) is difficult to measure with surface approaches. Our parsed corpora allow us to approximate these by retrieving singular nouns without a determiner and excluding NPs headed by a proper name (e.g. *I like Peter*). Simply sorting these data by frequency results in a list of generally frequent nouns rather than non-counts. Factoring in the probability of bare vs. non-bare NP is problematic as well because of data sparseness. Ranking by collocational force works considerably better. Typically, the significance-based T-score performs better than the information-theoretic Observed divided by Expected (O/E) or mutual information (MI) metric on this task. For an overview of collocation statistics, see Evert (2009). T-score is defined as (O-E)/√O. O are the corpus counts, E the co-occurrence frequency if words are randomly shuffled.

Table 2, which is sorted by descending T-score of zero-determiner + noun (column 2) lists the findings for the top 15 bare NP candidates; for the top 200 candidates combined, precision is 60%.[10] Precision describes the fraction of nouns that are true positives. At rank 5, for example, four out of the five nouns seen in the list from the top until here, are true positives, precision is thus 4/5 = 80%.

Table 2: Top candidates of *zero* + noun collocation, from BNC.

Noun	T.zero	f(noun. SING)	Manual Verdict	Rank	Precision T.zero
someone	0.9765	15,195	Y	1	1
something	0.9729	45,755	Y	2	1
none	0.9703	8682	Y	3	1
yeah	0.9160	7745	N	4	0.75
access	0.8995	8203	Y	5	0.8
mathematics	0.8955	895	Y	6	0.8333
bargaining	0.8945	559	Y	7	0.8571
cancer	0.8914	2865	Y	8	0.875
travel	0.8792	2211	N	9	0.7777
ft	0.8775	1117	N	10	0.7
alcohol	0.8769	2063	Y	11	0.7272
tobacco	0.8742	620	Y	12	0.75
km	0.8713	1186	N	13	0.6923
finance	0.8667	2446	Y	14	0.7142
discrimination	0.8602	1702	Y	15	0.7333

10 Note that the tagger treats *someone, something and none* as nouns.

A further aspect of property 1 is that non-counts are usually unmarked for number. With a precision of 70% for the top 200 candidates, this morphological property of nouns is the best single feature for the retrieval of potential non-counts. Figure 1 shows the cumulative precision (vertical axis) by rank (horizontal axis). At the rightmost position (250), the precision of the candidates is still almost 70%, with 174 of the 250 top candidates being true positives, and the curve only falls slowly. The fact that the curve falls, i.e. precision is highest to the left, indicates that the feature (the T-score of zero determiner plus noun) has a strong positive correlation to non-countability, which is a further indication that we use a meaningful operationalization and can also be interpreted as a cognitive signal: absence of a determiner prepares listeners for a non-count noun.

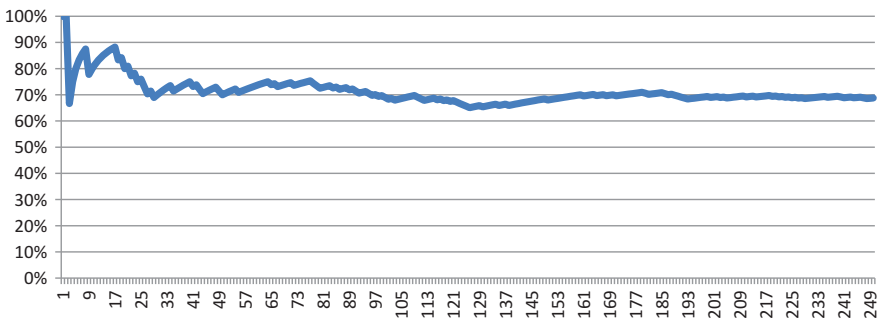

Figure 1: Precision of T-score collocation for singular + noun for the top 250 candidates (last column in Table 2).

Property 2: We approximate Krifka's "numerative constructions" by exploiting the fact that non-counts like *bread* often occur inside an *of*-PP construction modifying an NP headed by a measurement noun, e.g. *slice of bread*. Manual post-editing of the initial results yields the corpus-derived inventory of quantifier nouns in Table 3.

Once the collocation value obtained by *noun1 of noun2* is multiplied by a boosting factor (empirically set to 4) if *noun1* is in the quantifier noun inventory, our approximation of property 2 performs much better. The most highly ranked 20 candidates are given in Table 4.

Property 3: In an initial approach, we had also tried to use co-occurrence with *some*, but this proved significantly less successful than any of the other measures.

Table 3: Semi-automatically obtained inventory of quantifier nouns.

Quantifier nouns
number, range, lack, lot, amount, piece, group, bit, edge, example, bottle, glass, quantity, model, unit, row, item, hand, acre, stretch, pint, page, mile, pile, period, degree, copy, share, plenty, quarter, half, charge, round, volume, moment, body, word, glass, drink, amount, supply, jug, drop, cup, bowl, tin, litre, carton, slice, loaf, plate, chunk, hunk, basket, pound, slab, ounce, mug, pot, tray, flask, sip, gulp

Table 4: Top 20 of *noun1 of noun2* T-score (boosted retrieval).

Rank	T.noun1 *of* noun2	noun2	Manual
1	6.3241	Sherry	Y
2	6.0833	Tea	Y
3	5.8518	Silver	Y
4	4.8786	Toast	Y
5	4.5638	Brandy	Y
6	4.2387	vacancy	N
7	3.8896	Gin	Y
8	3.7194	Soup	Y
9	3.5121	Clarity	Y
10	3.4759	Rat	N
11	3.4169	Cocoa	Y
12	3.3721	Ham	Y
13	3.3121	academic	N
14	3.2349	Bean	N
15	3.1496	Bread	Y
16	3.1420	paperwork	Y
17	3.1180	Coffee	Y
18	3.0945	brochure	N
19	3.0882	Fun	Y
20	3.0755	champagne	Y

We noticed that rare words are hardly ever non-count and therefore introduced raw frequency of the noun as an additional feature. A linear combination of all our five features obtains a precision rate of 80% for the top 100 and 67% for the top 400 candidates extracted from the BNC. This list was then manually post-edited to remove all count nouns, which yielded a final list of 266 validated potential non-count nouns.

3.2 Evaluating the performance of combined collocation measures

In a next step, we linearly combine our four successful T-score features (*some* + noun (T.*some*), zero article (T.zero), prequalifier (T.*of*-PP), and singular (T.sing)), expecting that this should yield more promising results than the use of individual features. The top 20 candidates are listed in Table 5.

Table 5: Top 30 of weighted T-score features.

T.Combo 4 T-score	Noun	T.*of*-PP	T.some	OE.some	T.zero	T.sing	f(noun.SING)	Manual
948.36	tea	6.08	14.22	11.60	0.71	21.70	6369	y
649.47	money	1.67	21.67	4.98	0.59	50.83	32,483	y
400.00	information	1.21	13.80	3.53	0.68	51.02	30,296	y
384.98	evidence	1.29	19.23	6.01	0.61	41.92	20,569	y
343.37	coffee	3.12	14.09	12.66	0.64	19.08	4593	y
331.02	instance	2.73	13.04	8.55	0.78	15.42	8793	n
299.25	fun	3.09	8.60	9.20	0.78	18.75	4089	y
282.57	time	0.65	52.23	4.08	0.36	65.92	141,933	y
261.15	advice	1.92	10.55	5.45	0.68	27.50	8813	y
257.68	bread	3.15	10.53	13.09	0.71	15.61	3025	y
238.53	help	1.50	13.78	5.92	0.61	30.59	10,940	y
224.38	example	1.27	13.49	2.91	0.68	28.23	34,369	n
218.49	milk	2.96	8.81	9.48	0.71	16.44	3237	y
170.55	support	0.76	11.10	3.71	0.71	39.62	19,585	y
155.35	attention	2.16	5.40	2.09	0.65	31.68	12,927	y
150.67	m	1.35	7.42	6.06	0.82	22.13	6557	n
148.51	water	1.32	8.69	2.26	0.62	34.18	23,194	y
146.16	cash	1.97	6.99	5.49	0.73	19.72	4534	y
141.13	food	1.18	12.38	4.93	0.68	20.81	12,151	y
125.24	something	1.07	1.96	1.69	0.97	62.73	45,755	y

The precision for the combined retrieval approach is as follows: 90% for the top 50 candidates, 80% for the top 100, 77% for the top 200, 67% for the top 400, and 56.8% for the top 500. The fact that precision in the lower range of the list is still > 50%, only tailing off slowly, indicates that the list of non-count nouns is open.[11]

11 The performance of a system which makes a random choice (sometimes called baseline system) would be around 10–25% in the data of Baldwin and Bond (2003). The tail can be expected to converge towards this random performance.

In a next step, we tested the performance of features, both individually and in various combinations, using logistic regression to predict the count/non-count distinction of the first 500 items (see Table 6). Regression uses optimal feature weights instead of equal weight for reach feature (as e.g. Naïve Bayes does). The weights are also learnt from the data.

Table 6: Regression analysis for factors predicting non-count nouns.

```
> mass_aov < - aov (Bin.Dec ~ T.sing + T.zero + T.some + T.of PP + N count,
  data = mass_tscore, family = binomial)
> summary(mass_aov)
                Df    Sum Sq   Mean Sq   F value     Pr( > F)
T.sing           1      9.00     9.005    49.992   5.30e-12  ***
T.zero           1     16.99    16.994    94.347    < 2e-16  ***
T.some           1      0.01     0.008     0.045      0.832
T.of PP          1      4.31     4.310    23.927   1.36e-06  ***
Ncount           1      3.39     3.391    18.829   1.73e-05  ***
Residuals      494     88.98     0.180
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> mass_confmat = table (round (predict (mass_aov)),mass_tscore$Bin.Dec);
> mass_confmat;
     0   1
  0  132  48
  1   84 236
> precision = (mass_confmat [1,1] + mass_confmat [2,2]) / sum
  (mass_confmat);
> precision;
[1] 0.736
```

With the exception of *some* + noun, all collocation measures proved significant, and the ranking order is *zero-article* > *singular* > *quantifier* > *noun frequency*. Due to the optimized weighting of features in combination (and due to noun frequency as a significant factor), precision goes up from 56.8% (linear model) to 73.6% (regression model). After removing insignificant *some* + noun, precision increases further to 73.8% and if all interactions are included, precision is 76.0%.[12]

12 Only the interactions between *noun frequency* and *singular* and between *noun frequency* and *quantifier* are significant, so the independence assumption is not grossly violated, and the feature weights and their ranking can be interpreted confidently.

Compared to Baldwin and Bond (2003), our semi-automatic method is a compromise, using fewer, linguistically motivated features, paired with limited manual intervention to reach high precision, collocation statistics instead of Bayesian statistics. We are thus able to evaluate the importance of individual features.

3.3 Evaluation of the corpus-derived retrieval list

We scrutinized the list of 266 lexemes derived in our bottom-up approach from the BNC before using it to retrieve pluralized non-counts from the ICE corpora. As pointed out in 2.1, some lexemes are polysemous; *credits* (short for *credit points* in university contexts or other institutional contexts), *crickets*, *controls*, *redundancies*, *respects* and *supports* were exclusively used in the grammaticalized count sense of the word and thus excluded from the list. Nouns that are always used in their plural form (such as *mathematics* and *species*) were also excluded as they are never instances of extended pluralization. We also excluded nouns that regularly pluralize, such as *action*, *disagreement*, *scrap*, *space*, *similarity*, *time*, *truth* as well as irregular *life* and *leaf*. Finally, because of frequent tagging errors (inflected verb form rather than pluralized noun), we removed *works* and *helps* from our list of lexemes. The resulting list consisted of 241 potential non-counts, of which 154 types were attested in our ICE data, including most of the nouns used in Mohr's (2016) top-down approach (see also 4.2 for a more detailed comparison).

4 Results and discussion

4.1 Overall frequencies

Figure 2 gives the normalized frequencies for the 154 potential pluralized non-count types found in ICE.[13] While the ENL varieties all yield low-frequencies of these nouns, there is no clear categorical distinction into ENL and ESL but rather a gradient. The differences across all varieties are highly significant (chi-square contingency table test using raw counts of Figure 2, i.e. 9 varieties x candidates vs. word count, p = 4.2E-30 at df = 8), which means that this noisy data already delivers a reliable trend without human intervention, but the pair-wise test of

13 We normalize because the size of different ICE corpora varies slightly, with ICE-NZ, for instance, amounting to almost 1.2 million words instead of the 1 million-word target, also with official word counts often above 2000 words (Vine 1999: 26–65).

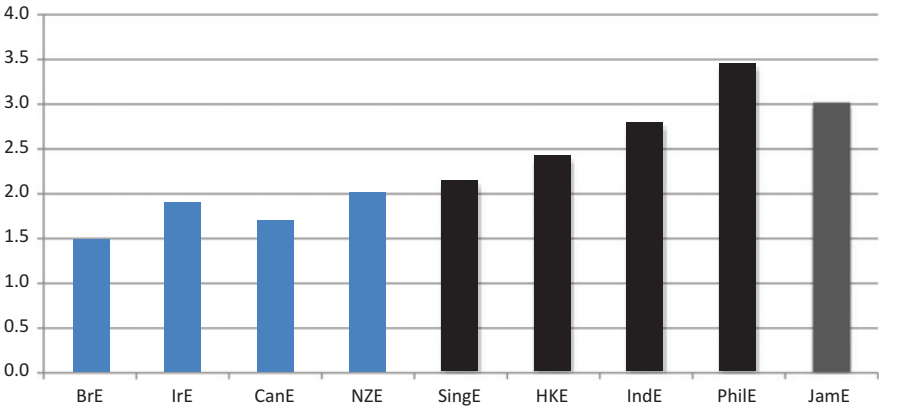

Figure 2: Relative frequency per 10,000 words of candidates for pluralized non-counts across WE (ICE).

the borderline varieties (NZE and SingE) is not (same chi-square contingency test, 2 varieties x candidates vs. word count $p = 0.10$ at $df = 2$). Note, however, that Figure 2 still contains instances of coerced count or type-noun readings and regularly pluralized forms of polysemous nouns. Coercion (differentiating between classes) and polysemy (technical terms as plural forms) is very frequent, particularly in scientific genres. Breaking down the results by mode reveals that the candidates are twice as frequent in writing than in speech.[14] Literary styles (monologue and printed) have higher frequencies than colloquial styles (dialogue and non-printed).

4.2 Results by lexeme

In a next step we broke down the totals by lexemes (sorted by decreasing frequency) in order to single out the quantitatively most salient contributors**.** Tables 7 and 8 give an overview of the most frequent types in ENL and ESL varieties, respectively.[15]

14 We should weight this against the fact that plural nouns are generally about 50% more frequent in writing in ICE. Pluralization is generally a feature of literary styles, at 20–50% more plurals than oral styles. Note that scientific reasoning is typically concerned with generalizing across types of objects, while spontaneous speech is often concerned with the individual. There are no discernable regional patterns in this trend.

15 Frequencies of items include capitalised variants.

Table 7: Most frequent plurals by lexeme (ENL).

ICE-GB	#	ICE-IRE	#	ICE-CAN	#	ICE-NZ	#
pressures	29	waters	25	pressures	21	waters	27
waters	22	pressures	18	waters	21	pressures	22
woods	9	foods	13	woods	19	foods	18
transports	8	cakes	12	foods	15	fruits	18
reliefs	7	precisions	10	convergences	10	cakes	11
wines	6	sweets	10	freedoms	8	sweets	9
cancers	5	chocolates	9	jurisdictions	6	protections	8
enthusiasms	4	wines	9	inputs	5	plastics	6
freedoms	4	woods	9	justices	5	woods	6
fruits	4	sugars	8	beers	4	coals	5
inputs	4	freedoms	4	fashions	4	oils	5
cakes	3	fruits	4	fruits	4	beers	4
cheeses	3	oils	4	cakes	3	fictions	4
foods	3	animations	3	comforts	3	injustices	4
medicines	3	cancers	3	injustices	3	overlaps	4
sleeps	3	inputs	3	rains	3	understandings	4
timbers	3	puddings	3	sensitivities	3	wines	4
alcohols	2	salts	3	soups	3	breakfasts	3
chocolates	2	silks	3	standings	3	cares	3
comforts	2	soaps	3	teas	3	freedoms	3
salts	2	teas	3	cancers	2	inputs	3
sweets	2	attentions	2	cares	2	salts	3
advices	1	breakfasts	2	chemistries	2	timbers	3
attentions	1	golds	2	coffees	2	coffees	2
breakfasts	1	meats	2	feedbacks	2	comforts	2
cares	1	medicines	2	golds	2	creams	2
cautions	1	persuasions	2	intelligences	2	gravels	2
clays	1	reassurances	2	praises	2	ironies	2

The highlighted plurals in Table 8 are lexemes that are missing from the most frequently attested ENL data. Interestingly, a number of these are shared across the ESL varieties. Importantly for our methodology, our retrieval missed very few data that are included in Mohr's (2016) top-down approach, namely *cattles* (3 instances in ICE-IND) and *offsprings* (9 instances, 1 from ICE-IND, 2 from ICE-PHI and 3 each from ICE-JAM and ICE-SING). Of the two potential mass nouns from her list that regularly pluralize and that were not included in our list (*stones* and *strings*), there are only four instances of *stones* in the ICE corpora (one each in ICE-HK and ICE-JAM and two in ICE-IND), which are not count nouns, such as (9):

Table 8: Most frequent plurals by lexeme (ESL/ESD).

ESL								ESD	
ICE-HK	**#**	**ICE-SING**	**#**	**ICE-IND**	**#**	**ICE-PHI**	**#**	**ICE-JAM**	**#**
waters	39	fruit	38	salts	53	rains	46	Waters	42
pressures	17	waters	26	fruits	27	waters	41	Fruits	37
cakes	16	**fishes**	**21**	medicines	21	fruits	29	Foods	27
clothings	**13**	cakes	19	rains	19	medicines	22	Inputs	19
fruits	13	pressures	17	waters	18	**flours**	**16**	Rains	18
informations	**13**	foods	11	pressures	17	**researches**	**12**	Oils	16
freedoms	10	rains	7	foods	15	pressures	11	Creams	11
medicines	8	salts	6	inputs	14	inputs	10	Sugars	11
fishes	**7**	cares	4	cancers	12	justices	9	Pressures	10
equipments	**6**	wines	4	sweets	12	**evidences**	**8**	**Musics**	**9**
foods	6	chocolates	3	**equipments**	**10**	**fishes**	**8**	Cakes	8
evidences	**5**	comforts	3	**fishes**	**8**	golds	8	Beers	7
homeworks	**5**	**equipments**	**3**	oils	8	soaps	7	Sweets	7
oils	5	golds	3	golds	7	**stuffs**	**7**	Meats	6
researches	**5**	inputs	3	freedoms	6	wines	7	Beer	7

understandings	5	terminologies	3	**percents**	**5**	**accommodations**	**6**	**Evidences**	**4**
beers	4	confirmations	2	plastics	5	plastics	6	Injustices	4
coffees	4	**gossips**	**2**	creams	4	**rices**	6	Powders	4
discriminations	4	inspirations	2	**furnitures**	**4**	silvers	6	Understandings	4
feedbacks	**4**	medicines	2	silks	4	woods	6	Cancers	3
inputs	4	oils	2	sugars	4	cakes	5	Comforts	3
plastics	4	plastics	2	cheeses	3	freedoms	5	**Fishes**	**3**
cancers	3	**researches**	**2**	coals	3	**homeworks**	**5**	Medicines	3
chocolates	3	soups	2	**informations**	**3**	overlaps	5	**Researches**	**3**
entertainments	3	woods	2	moneys	3	trainings	5	Silks	3
sweets	3	**accommodations**	**1**	chocolates	2	oils	4	Woods	3
advices	2	ammunitions	1	comforts	2	soups	4	Dissatisfactions	2
employments	2	animations	1	**enthusiasms**	**2**	discomforts	3	**Equipments**	**2**

(9) Old *stones* walls in Hong Kong have become a threatened heritage. (ICE-HK W2a-022)

Note that Table 8 shows the Zipfian distribution typical of lexical data. The top three types cover at least a third of all tokens in all varieties in Table 8. In addition, the top types reveal a certain regional bias likely to be due to local geographic and climatic conditions such as torrential rains and coastal waters. As *rains* and *waters* are given type-of readings in these contexts, qualitative analysis is necessary to distinguish between regular and non-standard extension of pluralization of non-counts.

4.3 Results by token

We annotated the resulting concordance (2,312 tokens) for (a) false positives, (b) regular plurals (both polysemous and non-polysemous), (c) coerced type-noun and plural uses and (d) extended (non-standard) plurals. Examples of false positives are genitive *today's* without the apostrophe or nouns in fixed phrases which are always pluralized, as in *all intents and purposes* (see (10)). Also excluded as false positives were tagging errors (e.g. (11)) and two instances of object language use of non-standard pluralized non-counts (see (12)).[16]

(10) And as a result of that then again for all *intents* and purposes the inertial force of the cell is virtually zero. (ICE-GB S2A-051)
(11) Nobody *cares* about the cat. (ICE-CAN A1A 073)
(12) Typical examples include ... plural form of an uncountable noun (e.g. *we need more *informations* and *equipments*). (ICEHK W2A 006)

16 In our qualitative analysis of the concordances we frequently needed to consider the larger context in order to decide whether we were dealing with a coerced type-noun reading or a non-standard extension of the noun. A serendipitous discovery of a tagging error (i.e. where the noun *gossips* had been tagged as verb (see below) shows that retrieval of pluralized non-counts from annotated corpora will also occasionally result in lower recall for individual lexical items than a top-down retrieval from an orthographic corpus would do.

A: All the gossips

B: Ya gossips very good (ICE-SIN S1A-046)

Out of 8 instances of *gossips* in our data, 6 were correctly identified as plural nouns and one correctly as a proper noun (the title of a picture).

Examples of regular plurals are given in (13)–(16).[17] Coerced count and type-noun readings are illustrated in (17)–(19), and genuine extended non-standard uses in (20)–(22).

(13) Justice Lewis F. Powell cast the decisive vote breaking a four-four deadlock among the eight other *justices*. (ICE-IND W2B-011)
(14) But unfortunately the government was too busy ... in the restructuring of the *securities* and *futures* market. (ICE-HK S2B-035)
(15) It had been hoped India would win seven or eight *golds* in a medal tally of around 45. (ICE-IND W2E-003)
(16) okay my bouquet to the *coppers*[18] (ICE-NZ S1B-038)
(17) The *salts* are also secreted through Gycathode by the process of guttation. (ICE-IND W1A-014)
(18) Consideration has to be given to the fact that the concentration of free *sugars* (i.e. monosaccharides) is increased as fruits ripen. (ICE-JAM W2A-030)
(19) Fruit vendors have many colourful citrus *fruits* for the first weeks of Lunar New Year: oranges, tangerines and kumquats chief among them. (ICE-HK W2D-011)
(20) Occasionally, large schools of tiny silvery *fishes* move around quickly, probably fleeing away from predators. (ICE-HK W2D-017)
(21) ... I will send more *informations* on this. (ICE-HK W1B-009)
(22) Forget about those *homeworks* and exam. (ICE-HK S1A-093)

Table 9 reports the distribution of regular plurals, coerced and extended uses in our manually annotated ICE data. In distinguishing these different contexts for pluralisation of non-counts, we go significantly beyond previous research (notably the only previous bottom-up study by Schmidtke and Kuperman 2017). The figures reveal that the majority of the plurals turn out to be regular uses. This is due to the large number of polysemous nouns. The frequency of extended non-counts per 10,000 words is given in Figure 3. Now, also pair-wise comparison of the differences between each ENL and ESL variety are highly significant. The proportion of extended uses (against coerced non-counts) across WE is shown in Figure 4.

The results confirm that extended uses of plural non-counts are more frequent in ESL/ESD than in ENL varieties. Interestingly, our analysis shows that HKE and SingE – the varieties with similar substrates – yield similar relative frequencies but not comparable proportions of extended pluralized non-counts,

17 We include metonymical uses of mass nouns such as *golds* for *gold medals* among the polysemous instances rather than coerced type-noun readings.
18 The context (testing for drunk driving) makes it clear that the noun refers to policemen.

Table 9: Frequencies of regular, coerced and extended (non-standard) plurals in ICE.

	regular plural	coerced non-count	extended non-count	total
ICE-CAN	146	54	3	203
ICE-GB	113	38	2	153
ICE-HK	122	70	62	254
ICE-IND	118	167	43	328
ICE-IRE	111	85	4	200
ICE-JAM	119	161	28	308
ICE-NZ	121	99	4	224
ICE-PHI	172	175	43	390
ICE-SING	86	107	26	219

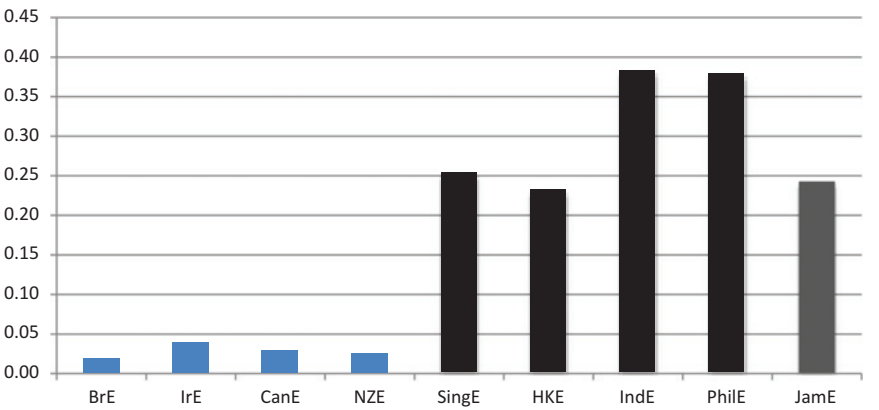

Figure 3: Relative frequency of extended non-counts (per 10,000 words).

indicating that substrate influence is unlikely to be the sole explanatory factor for extension of plural morphology to non-counts. IndE, PhilE and SingE are similar in their use of extended plurals, as are the ENL varieties.

Extended uses come from a total of 58 types (i.e. more than twice as many as the number of types included in Mohr's exclusively top-down study). Examples of extended plurals that have not previously been reported in research on ESL varieties are for instance *attentions*, *appreciations*, *bloods*, *funs*, *fundings* and *nonsenses*. Crucially, these extended uses are not limited to (spontaneous) spoken contexts (e.g. (25)) but are also regularly attested from formal written material, as in (23) and (27).

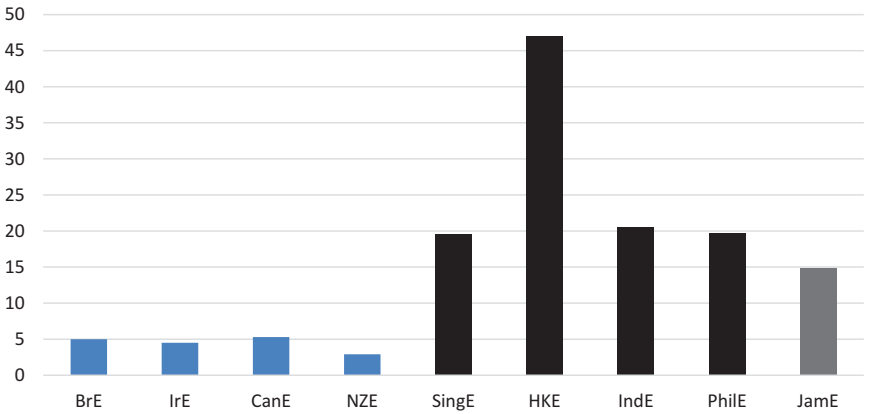

Figure 4: Relative frequency (%) of extended non-counts (against coerced plurals).
Note: Note that the proportional frequencies have to be taken with a pinch of salt as there may be a problem with burstiness skewing the data. In the case of *flours*, a single text (ICE-PHI W2A 039) was the source of the coerced reading of a mass noun.

(23) But laws ... have been slow coming, partly because of inadequate *appreciations* of environmental problems ... . (ICE-IND W1A 009)
(24) Black driplets of *bloods* were frozen around his nostrils .... (ICE-IND W2F 018)
(25) ... she have to have *funs* very active and somewhat like that. (ICE-HK S1A 029)
(26) I hope this was not just a one-time investment in computers with not enough follow-up *fundings* ... . (ICE-HK S2B 021)
(27) The Gallic concerns with things ontological were part of the apparatus of assimilationist education meted out to French-speaking Africans to whom the search for *essences* came more naturally than to their Anglophone counterparts, whose Anglo-Saxon education often eschews such essays into what Edmund Burke, the English philosopher, might well have dismissed as metaphysical *nonsenses*. (ICE-JAM W2A 013)[19]

Interestingly, non-standard plurals in ESL have even made their way into idioms, here the light verb *take care*[20]:

19 Note that the example yields a serendipitous find of an extended pluralisation (i.e. *essences*) not retrieved by our bottom-up approach.
20 An alternative analysis would be that speakers of HKE treat *take care* as a compound VP with an inflectional ending for person attached rather than plural morphology. Under this analysis, we would be dealing with verbal group inflection rather than extended noun pluralization.

(28) Because I think he *take cares* of me in all aspects as well. (ICE-HK S1A-018)

Contrary to popular belief, extended uses of non-count pluralization are also attested in ENL corpora[21]:

(29) There's never any blaming game that goes on in the teacher's *knowledges* … . (ICE-CAN S2A 045)
(30) "You boys want *coffees*?" (ICE-CAN W2F-007)
(31) The fleshy *fruits* of Coffea arabica contain two "beans" each … . (ICE-NZ W2B 025)
(32) The Managements of the three schools and representatives of all the *staffs* met then with the Independent Facilitator … . (ICE-IRE S2C 016)

Classification of individual instances may be problematic: the plural in (31), for instance, could be said to be an instance of a coerced count use licensed by the pronoun *each* in the immediate context rather than a pluralized non-count of *fruit*. Similarly, *staffs* in (32) could be argued to refer to the three bodies of employees from the different schools, i.e. also be an instance of a contextually motivated pluralization rather than a non-standard use. Distinguishing between extended and polysemous uses is also often difficult. Example (33) might, at first sight, appear to be a non-standard, extended use of *informations* in an ENL variety:

(33) … that's the telescope used to obtain these *informations*. (ICE-GB S2A-058)

According to OED, the noun has a ("rare") sense that is a count-noun, namely "a fact or circumstance of which a person is told; a piece of news or intelligence". Interestingly, the last attested example (1959) in OED, like the ICE-GB occurrence, is also from a scientific context:

(34) Scientific prediction, in contrast with prophecy, is based on laws and on specific reliable *informations* regarding the present (or past) state of affairs. (OED, s.v. *information,* n. 2.b., 1959, M. Bunge Metasci. Queries ii. 52)

Against this evidence, it is difficult to argue that the use in (33) is really non-standard. Similarly, a noun that at first might appear as a good example of a prototypical non-count noun is actually one that has both non-count and count

21 Mohr (2016: 178–179) presents frequency information on her set of nouns in the BNC (her benchmark corpus) but only provides an example of a coerced type-noun use of *cheeses* rather than non-standard extensions of pluralization.

senses recorded in the OED (s.v. *research*, n. 2a. and 2b.), with the count uses attested regularly in the nineteenth century but probably less frequent in the twentieth (but see (35)). It is therefore difficult to confidently analyse instances of plural *researches* as extended uses. While (36) and (37) are plausible as a continuation of the older (lexicalized) count noun in that *researches* refer to individual studies, (38) is more likely to refer to the general need for research and was therefore analysed as an extended use of the non-count noun. The analysis is a matter of interpretation, however, and (39), which we took to refer to the author's *studies*, could also have been interpreted as a non-standard plural.

(35) His researches on the fossil woods led him to *researches* on other fossils. (OED, s.v. research, n. 2b., 2002, D. Freedberg *Eve of Lynx* iii. xi. 344,
(36) ... Campbell's commonsense aversion to historical speculation ... led him gradually to consider the avid *researches* of contemporary Irish antiquarians as having more in common with the fiction of Macpherson than with the dispassionate, polemic published in the changed circumstances of 1789, ... (ICE-IRE W2A-010)
(37) Okay for example one of his important *researches* is that of group decision making. (ICE-PHI S1B 006)(38)
(38) A review of literature showed the lack of current information on influenza in the Philippines and this all the more emphasizes the need to conduct new *researches* on the virus and its epidemiology in the local setting. (ICE-PHI W2A 023)
(39) Although I have always included language variables in my *researches*. (ICE-PHI, S1B 001)

In order to gauge the consistency of the annotation and effects of inter-annotator disagreement, we had a subset of 811 variable instances coded by a second annotator, a native speaker. The instances had been randomized so as to reduce the possible impact of priming, and information on the variety had been removed. It turned out that inter-annotator agreement was quite low, at 61%. The following examples may illustrate why this is the case: in each instance, only one annotator had given a (standard) type-of interpretation to *fruits*, either because the context was interpreted differently or there was not enough context to decide:

(40) Attractive spiny *fruits* and coloured leaves make some bidibids popular groundcover plants for gardens. (ICE-NZ W2B-025)[22]

22 The native speaker's argument was that "this is standard because the speaker is talking about different varieties of plants and therefore different varieties of fruits," whereas we would

(41) ... they would often come from *fruits* or vegetables and dyes but the ones that were to last really long would also be from rocks and minerals. (ICE-CAN S2A-030)
(42) Oh we have too many *fruits* you know. (ICE-SING S1A-064)

Similarly, a type-of reading and a general (count-noun) reading were given to *medicines* in the following example, resulting in divergent codings:

(43) The reasons for which we neglect taking our *medicines* as advised include forgetfulness fear of side-effects and misunderstandings of the instructions provided. (ICE-GB S2B-038)

The larger context may help to decide whether an individual instance can be given a coerced type-noun interpretation or a non-standard extension reading. Out of context, (44) might appear to be a non-standard extension of a non-count, but a look at the larger context opens the possibility of a type-noun reading as the recipe for *Busha Browne's Hearty Red Pea Soup* combines soup meat with bacon or a salted pig's tail, i.e. different kinds of meat.

(44) When peas and coco are cooked, remove the *meats*. (ICE-JAM W2D-012)

Elsewhere, the native speaker drew on orthography to decide whether a certain pluralized non-count was a standard use, claiming that "*foodstuffs* as one word is standard, *food stuffs* as two is not," thus making an extended non-count. For (46), one annotator suggested coercion to types of fast food whereas the other argued coercion of an element in a list was less likely than analogy with the other nouns in the conjoined NP (resulting in a non-standard plural).

(45) Urban dwellers ... have traditionally had their basic food *stuffs* subsidized. (ICE-CAN S2B-035)
(46) Neither except most perfunctorily does it show their reaction to waking up in a world of TV junk *foods* miniskirts nuclear weapons rock and roll and all the other wondrous gifts of modern civilisation. (ICE-GB S2B-033)

have argued that the different varieties of the plant all have the same kind of spiny fruit (i.e. not different spiny fruits), making the plural non-standard. *Bidibid* is a New Zealand word for a type of evergreen creeping plant, derived from Māori *piripiri*.

Interestingly, the number of plurals from ENL varieties that were coded as nonstandard by the native speaker was considerably larger (40:13) than those we had coded as non-standard in the same data set (including, among others, examples (42) and (51)). Overall, the native speaker annotator rated more plurals as non-standard, as Figure 5 shows. Annotator 1 is one of the authors, Annotator 2 a native speaker.

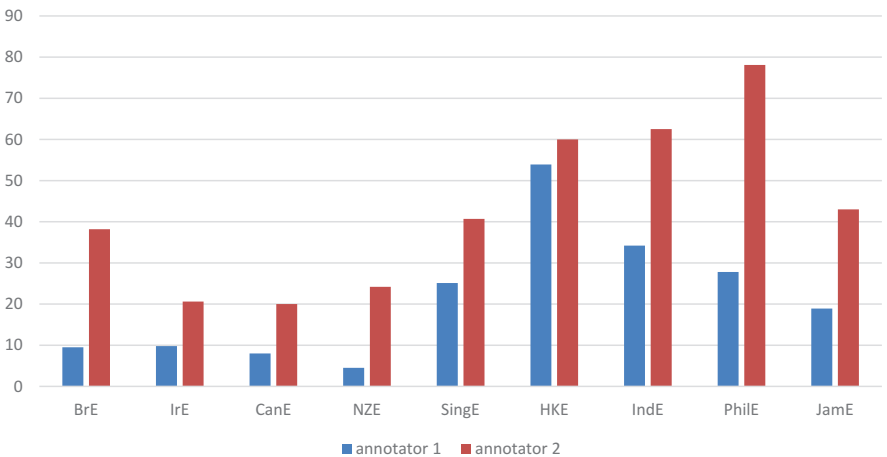

Figure 5: Distribution of extended non-counts (as % of coerced and extended uses) across varieties by annotator.

The fact that inter-annotator agreement is quite low is a result of the gradient nature of countability in English, but also of the role that context plays in the interpretation of examples. Some divergent ratings might well have to be attributed to the fact that we cross-validated all our ratings against OED dictionary entries and erred on the side of caution whenever a coerced reading might have been possible, while the native speaker was less conservative. For individual nouns, moreover, native speaker evaluations vary, as is the case with *feedback(s)*.[23] Importantly for our initial hypotheses, the two ratings confirm a general divide into ENL vs. ESL/ESD varieties as a tendency, with less of a difference between BrE and SingE in the native-speaker ratings than might be expected.

23 See https://www.englishforums.com/English/FeedbackOrFeedbacks/brdkr/post.htm or https://www.englishforums.com/English/PluralForFeedback/vlrx/post.htm

5 Summary and conclusion

Our bottom-up, theory-informed approach to derive potential non-count nouns from corpora on distributional grounds works well: it detects some 90% in the top 50 items, and up to 76% in the top 500 items. Moreover, by combining a bottom-up corpus-based approach with fine-grained qualitative analyses we can provide a more nuanced view of pluralisation of non-counts across ENL and ESL. On a cognitive and usage-based level, we have learnt that the count/noun-count distinction can largely be learnt from word distributions (Klein and Manning 2001; Mintz et al. 2014). Importantly, our method misses very few potentially non-standard pluralized non-counts (Section 4.2). Instead, it allows us to report instances of less prototypical non-standard pluralization (such as *attentions, appreciations, bloods, funs, fundings* and *nonsenses*) that remained under the radar in previous top-down studies (Section 4.3). Despite noisy data, there are clear quantitative differences between ENL and ESL varieties, the latter having a higher tendency for non-count noun pluralization. The overall picture is more complicated, though, and borders between variety types are not clear-cut, as ESL varieties show quantitative differences and do not form a coherent group. It is in this context that the different approaches to normalization have real consequences for the interpretation of the data: it is only by looking at the proportion of extended pluralization (against coerced pluralization) that we see a difference between HKE and SingE emerging which goes beyond a simple explanation in terms of "substrate influence". Qualitative analyses are thus crucial if one wants to move beyond simple frequency-based explanation.

Our subsequent qualitative analyses reveal that the majority of pluralized non-counts in both ENL and ESL varieties are coerced type-noun instances. In other words, simply retrieving non-counts from corpora is not enough to argue for an extended (non-standard) use of the category. The analyses of our ICE data further show that, contrary to the widespread assumption that non-standard pluralization of nouns like *furniture* or *information* are exclusively found in second language varieties, these are also occasionally attested in native speaker varieties. We attribute this to the gradient rather than categorical distinction between count and non-count nouns. This, alongside the availability of contextual information, also proved a challenge for the qualitative analysis and became evident in the inter-annotator disagreement. While studies hinging on this criterion need to be assessed critically, low inter-annotator agreement crucially did not impact the observation of a general divide between ENL and ESL. Moreover, our findings confirm Denison's (1998: 98) observation (based on a 1993 Linguist List discussion) that sporadic use of non-count nouns as count

nouns (e.g. *homeworks*[24] and *surgeries*) is possible in BrE and AmE. However, whether this is a recent trend led by AmE (among the ENL varieties) needs further empirical support.

Finally, the fact that we find non-standard extension of pluralization to non-counts in ENL varieties is of relevance for theories of origin as well. Most previous research explains extended pluralization of non-counts as arising from processes of second-language acquisition and contact-induced hypercorrection. Such a view rests on the assumption that the (British English) varieties that served as inputs throughout the currently English-speaking world did not have the feature, for which we provide counter-evidence. Alternatively, our findings suggest that the feature was present in the superstrate. In other words, it is not enough to resort to substrate influence or L2-acquisition processes as an explanation for apparently "nativized" pluralized non-counts such as *researches*. Data from the Old Bailey court proceedings, in particular, show that these "extended" uses of plural non-counts are also regularly attested in earlier stages of BrE and are thus likely to have been part of the input varieties that helped form the ESL varieties. Future research should ideally be able to draw on historical ESL corpora to verify the continuity of this feature across time. Including singular instances of non-count nouns in future research will allow us to model predictor variables for the use of regular and extended pluralized non-counts. It would also be useful to compare the ESL to EFL corpus data to further confirm Hall et al.'s (2013) finding that non-institutionalized varieties pattern more closely with ENL in this area of grammar.

References

Corpora
BNC = *British National Corpus*. via Depedency Bank, see Lehmann and Schneider 2012.
COHA = *Corpus of Historical American English*. https://corpus.byu.edu/coha/.
COOEE = *Corpus of Oz Early English*. http://www.helsinki.fi/varieng/CoRD/corpora/COOEE.
ICE = *International Corpus of* English. http://www.ice-corpora.uzh.ch.
Old Bailey. https://www.oldbaileyonline.org (accessed 10.11.2016).
Secondary sources
Allan, Keith. 1980. Nouns and countability. *Language* 56(3). 541–567.

24 An example from a Language Log post by Mark Liberman (February 6, 2010) shows that *homeworks* is possible in AmE: "And in order to prevent some students from relying on the archives of past *homeworks* and exams stored (I'm told) at some fraternities and sororities, I need to find new examples every year" (http://languagelog.ldc.upenn.edu/nll/?p=2100).

Alo, Moses A. & Rajend Mesthrie. 2004. Nigerian English: Morphology and syntax. In Bernd Kortmann, Edgar Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A handbook of varieties of English*, vol. 2. 813–827. Berlin: de Gruyter.

Baldwin, Timothy & Francis Bond. 2003. Learning the countability of English nouns from corpus data. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1* (ACL '03), vol. 1, 463–470. PA, USA: Association for Computational Linguistics, Stroudsburg.

Cruse, D. Alan. 1999. Number and number systems. In Keith Brown & Jim Miller (eds.), *Concise encyclopedia of grammatical categories*, 267–271. Oxford: Pergamon.

Davies, Mark & Robert Fuchs. 2014. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1). 1–28.

Denison, David. 1998. Syntax. In Susanne Romaine (ed.), *The Cambridge history of the English language, Vol IV 1776–1997*, 92–329. Cambridge: Cambridge University Press.

Deshors, Sandra, Sandra Götz & Samantha Laporte. 2016. Linguistic innovations in EFL and ESL: Rethinking the linguistic creativity of non-native English speakers. *International Journal of Learner Corpus Research* 2(2). 131–150.

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 1212–1248. Berlin: de Gruyter.

Grimm, Scott & Beth Levin. 2011. *Between count and mass: Furniture and other functional collectives*. Stanford University. http://web.stanford.edu/~bclevin/lsa11talk.pdf (accessed 19 January 2015).

Grimm, Scott & Beth Levin. 2012. *Who has more furniture? An exploration of the bases for comparison*. Universitat Pompeu Fabra and Stanford University. http://web.stanford.edu/~bclevin/paris12mcslides.pdf (accessed 19 January 2015).

Hall, Christopher J., Daniel Schmidtke & Jamie Vickers. 2013. Countability in world Englishes. *World Englishes* 37(1). 1–22.

Harris, Zellig. 1954. Distributional structure. In J. A. Fodor & J. J. Katz (eds.), *The structure of language*, 33–49. Englewood Cliffs, N.J.: Prentice-Hall.

Hundt, Marianne. 2015. World Englishes. In Douglas Biber & Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics*, 381–400. Cambridge: Cambridge University Press.

Jackendoff, Ray. 1991. Parts and boundaries. *Cognition* 62(2). 169–200.

Joosten, Frank. 2003. Accounts of the count-mass distinction: A critical survey. *Nordlyd* 31(1). 216–229.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk & H. G. Widdowson (eds.), *English in the world: Teaching and learning the langauge and literatures*, 11–30. Cambridge: Cambridge University Press.

Klein, Dan & Christopher Manning. 2001. Distributional phrase structure induction. *Proceedings of the 2001 Workshop on Computational Natural Language Learning* 7: 14:1–14: 8.

Kortmann, Bernd & Kerstin Lunkenheimer, eds. 2011. *The electronic world Atlas of varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://www.ewave-atlas.org/ (accessed 12 January 2018).

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2012. *The mouton world atlas of variation in English*. Berlin: de Gruyter.

Krifka, Manfred. 1999. Mass expressions. In Keith Brown & Jim Miller (eds.), *Concise encyclopedia of grammatical categories*, 221–223. Oxford: Pergamon.

Lehmann, Hans Martin & Gerold Schneider. 2012. BNC Dependency Bank 1.0. In Signe Oksefjell Ebeling, Jarle Ebeling & Hilde Hasselgård (eds.), *Studies in variation, contacts and change in English, Volume 12: Aspects of corpus linguistics: compilation, annotation, analysis*. Helsinki: Varieng. http://www.helsinki.fi/varieng/journal/volumes/12/ (accessed 01 January 2018).

Mair, Christian. 2017. Crisis of the outer circle? – Globalisation, the weak nation state, and the need for new taxonomies in World Englishes research. In Markku Filppula, Anna Mauranen, Juhani Klemola & Svetlana Vetchinnikova (eds.), *Changing English: Global and local perspectives*, 5–24. Berlin: Mouton de Gruyter.

Meriläinen, Lea & Heli Paulasto. 2017. Embedded inversion as an angloversal: Evidence from inner, outer and expanding circle Englishes. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of world Englishes*, 676–696. Oxford and New York: Oxford University Press.

Mesthrie, Rajend. 2012. *Black South African English*. In Kortmann & Lunkenheimer (eds.), 493–500. Berlin: Mouton de Gruyter.

Mesthrie, Rajend. 2017. World Englishes, second language acquisition, and language contact. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of world Englishes*, 175–193. Oxford: Oxford University Press.

Mesthrie, Rajend & Rakesh M. Bhatt. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge: Cambridge University Press.

Mintz, Toben H., Felix Hao Wang & Vivian Jia Li. 2014. Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. *Cognitive Psychology* 75C. 1–27.

Mohr, Susanne. 2016. From Accra to Nairobi – The use of pluralized mass nouns in East and West African postcolonial Englishes. In Daniel Schmidt-Brücken, Susanne Schuster & Marina Wienberg (eds.), *Aspects of (Post)Colonial Linguistics*, 157–187. Berlin: de Gruyter.

Mukherjee, Joybrato & Marianne Hundt (eds). 2011. *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*. Amsterdam/Philadelphia: John Benjamins.

Oxford English Dictionary. http://www.oed.com.

Quirk, Randolph, Sidney Greenbaum, Goffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of English*. London: Longman.

Sand, Andrea. 2012. *Jamaican English*. In Kortmann & Lunkenheimer (eds.), 210–221. Berlin: Mouton de Gruyter.

Schmidtke, Daniel & Victor Kuperman. 2017. Mass counts in World Englishes: A corpus-linguistic study of noun countability in non-native varieties of English. *Corpus Linguistics and Linguistic Theory* 13(1). 135–164.

Schmied, Josef. 2008. East African English (Kenya, Uganda, Tanzania): Morphology and syntax. In Rajend R. Mesthrie (ed.), *Varieties of English. Africa, South and Southeast Asia*, 451–471. Berlin: de Gruyter.

Schmied, Josef. 2012. *Tanzanian English*. In Kortmann & Lunkenheimer (eds.), 454–463. Berlin: Mouton de Gruyter.

Schneider, Edgar W. 2015. Models of English in the world. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of world Englishes*, 35–57. Oxford: Oxford University Press.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. University of Zurich PhD Thesis.

Schneider, Gerold & Gaëtanelle Gilquin. 2016. Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research* 2(2). 177–204.
Sharma, Devyani. 2012. *Indian English*. In Kortmann & Lunkenheimer (eds.), 523–530. Berlin: Mouton de Gruyter.
Taiwo, Rotimi. 2012. *Nigerian English*. In Kortmann & Lunkenheimer (eds.), 410–416. Berlin: Mouton de Gruyter.
Tomasello, Michael. 2000. The item based nature of children's early syntactic development. *Trends in Cognitive Sciences* 4. 156–163.
Vine, Bernadette. 1999. *Guide to the New Zealand component of the international corpus of English*. School of Linguistics and Applied Language Studies, Victoria University of Wellington.
Wong May, L.-Y. 2012. *Hong Kong English*. In Kortmann & Lunkenheimer (eds.), 548–561. Berlin: Mouton de Gruyter.
Ziegeler, Debra. 2010. Count-mass coercion, and the perspective of time and variation. *Constructions and Frames* 2(1). 33–73.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cllt-2018-0068).

Bionotes

Gerold Schneider

Gerold Schneider is a computational linguist and corpus linguist. He is currently senior lecturer (Privatdozent) and postdoc staff member at the Institute of Computational Linguistics and permanent staff member (scientific programmer and senior lecturer) at the English Department of the University of Zurich. He studied English, Computational Linguistics and General Linguistics. He developed a large-scale syntactic parser for English as part of his doctoral degree. His habiliation deals with the connections between Text Mining, theoretical linguistics and psycholinguistics. He likes to apply computational linguistic methods to various linguistic fields and to Digital Humanities.

Marianne Hundt

Marianne Hundt is Professor of English Linguistics at Zürich University. Her research interests range from grammatical change in contemporary and late Modern English to varieties of English as a first and second language (New Zealand, British and American English; English in Fiji and South Asia) and language in the Indian Diaspora. She has been involved in various corpus compilation projects and is the co-coordinator of the International Corpus of English. She has also explored the use of the world-wide-web as a corpus and for corpus building. She is the author of New Zealand English Grammar, 1998, co-author of Change in Contemporary English, 2009 and co-editor of English World-Wide (since 2013).

Daniel Schreier

Daniel Schreier is Professor of English Linguistics at the University of Zurich. His research interests include varieties of English around the world and sociolinguistics. He is author of several books on English in the South Atlantic, has published some 60 articles and has served as co-editor of English World-Wide (2013–2019).